This is a preprint of an article accepted for publication in Cataloging & Classification Quarterly
(Lee, D. J., & Stvilia, B. (in press). Developing data identifier taxonomy, Cataloging & Classification Quarterly)

DEVELOPING DATA IDENTIFIER TAXONOMY

Dong Joon Lee & Besiki Stvilia

ABSTRACT

As the amount of research data management is growing, the use of identity metadata for discovering, linking and citing research data is growing too. To support the awareness of different identifier systems and comparison and selection of identifier for a particular data management environment, there is need for a knowledge base. This paper contributes towards that goal and analyzes the data management and related literatures to develop a data identifier taxonomy. The taxonomy includes four categories (domain, entity types, activities, and quality dimensions). In addition, the paper describes 14 identifiers referenced in the literature and analyzes them along the taxonomy.

Keywords: Identifier, research data, quality requirements

INTRODUCTION

Funding agencies now require applicants to submit plans for disseminating and providing access to research data.[1] In addition, many journals and article databases now require the submission of data along with manuscripts, as well as the annotation and integration of the manuscript's content with the data.[2] All of these requirements were intended to increase the access and use of research data. Access to research data used in the production of outcomes has become essential for understanding the research.[3] The need for greater access and sharing of research data to increase the impact and efficiency of scientific activities and funding has been emphasized by the government and various funding agencies.[4] Greater access to research data, however, is enabled not just by appropriate policies but also by the deployment of effective infrastructure mechanisms including augmenting data with effective metadata.[5] Identifiers are important metadata that traditionally have been used for entity identification, linking and referencing in various domains.[6] This paper examines identifier system use with research data. To enable effective metadata creation support for research data, it is essential to gain better understanding of the current uses of identifiers with research data as well as the needs for identifier system functionalities and the functionalities of currently available identifier systems.

PROBLEM STATEMENT

With increased push for data sharing and reuse by the government, funding agencies and scholarly communities, there is increased attention on the design of metadata for data, including identifier schemas.[7] As different communities manage different data on different entities, identifier schemas are contextual and tailored to the community's data practices.[8] In molecular biology, Life Science Identifiers (LSIDs) are used to identify and integrate data objects distributed in multiple databases.[9] In chemistry, chemical identifiers (e.g., Chemical Abstracts Service (CAS) Registry Number, International Chemical Identifier) and their associated metadata help discover chemical substances and compounds.[10] Large academic publishers have made important changes too. For example, Thomson Reuters announced its development of a data citation index and started indexing research data from repositories available to them across disciplines and around the world to supplement articles in the Web of Knowledge with associated research data. Robust identifier systems are essential for making these connections.[11] Likewise, Elsevier decided to use Open Researcher and Contributor ID (ORCID) to create more robust links between scholarly works and their authors.[12] These changes from diverse disciplines and major publishers reflect the increased uses and importance of identifiers in the current research environment.

Surprisingly, the practical use of identifier systems for research data and their activities has not yet been systematically studied in the literature. This paper examining the gap between identifier systems used by different communities and analyzing them along different facets of their design and use would be invaluable and could be used by data managers and curators as a knowledge tool in selecting an identifier system(s) for their data repositories. In addition, the paper develops a taxonomy of identifier system characteristics discussed in the literature which can be used by data managers and curators as a knowledge tool in selecting an identifier system(s) for their data repositories. The paper can also inform policy development for institutional data repositories with regard of identifiers schema selection and use.

DEFINITION OF DATA IDENTIFIER

Identifiers can be defined in many different ways depending on the purposes (e.g., identification, reference, annotation) for which they are applied. The Oxford English Dictionary defines identifier as "a

thing used to identify someone or something" or "a sequence of characters arbitrarily devised to identify or refer to a set of data, a location in a store, or a point in a program." This definition highlights the purposes of identifying and referencing objects. Altman and King,[13] who discussed a possible schema for data citations, characterized identifier as "a character string guaranteed to be unique among all such names, which permanently identifies the data set independent of its location." Their definition points to the importance of identifier systems' performance (e.g., persistent access) as well as the purposes of data entity disambiguation. Pepler and O'Neil,[14] in their definition of identifier, specified the resources (e.g., person, house, color, employee, journal paper, or file) referenced by the identifier. In a recent report from NISO/NFAIS,[15] a definition highlighting identifiers' overall purpose was offered: Identifiers "provide discoverability of and linking to content." Based to these definitions, we can conclude that the definition of identifiers should mention identifier system features, assigned entity types, and purposes of identifiers. The set of data entity types that need to be referenced by identifiers is contextual and varies from one discipline to another. Likewise, different identifier systems can be used for referencing different kinds of entities. However, the activities (purposes) of identifiers do not change much. For the purposes of this paper, and based on our literature analysis, we define a data identifier as a sequence of symbols designed to identify, cite, annotate, and/or link research data and their associated metadata.

METHODOLOGY

This paper surveys identifier systems used in research data management by different communities and analyzes them along the different facets of their design and use. These include the data entities and types of activity identifiers are used for, and conceptualizations of identifier quality. To complete this, we developed a data identifier taxonomy to help understand data identifiers, based on an extensive literature review. The study reviewed 70 different articles related to the issues of research data management and identifier schemas from both practice- and academic-focused journals. The article selection from both types of journals allowed reviewing practical systems and research issues. 40 different sources obtained from various institutional websites and conference presentations were also reviewed and they added additional contextual understandings on identifier practices. These issues and characteristics discussed in

all the sources then were organized into a taxonomy. The taxonomy consists of four different categories: domains, entity types, activities, and quality dimensions. Three different conceptual models designed for information resource organization (i.e., FRBR, PREMIS, and CIDOC CRM) guided the identification of the entities of data identifiers. The taxonomy can help librarians, repository managers, scholarly communities, system developers, and publishers in the selection or design identifier systems for their data repositories.

RESEARCH QUESTIONS

The paper used a literature analysis to addresses the following research questions:

What are the identifier systems used in different domains for different data entities?

What are the types of data entities identifiers are used for?

What are the types of activities, which use identifiers?

How is identifier quality perceived?

DATA IDENTIFIERS

Different communities use different identifiers for different data entities. Fourteen identifiers referenced by multiple articles in the literature and used for research data entities are selected and reviewed in this section. The brief descriptions of each identifier help move forward this study.

Archival Resource Key (ARK)

In 2001, Kunze and Rogers at the U.S. National Library of Medicine originally developed the ARK. It is currently maintained at the California Digital Library. The ARK, which is used to identify research data in institutional repositories, is a domain-independent identifier.[16] It enables users to access the metadata of the assigned object. The identifier is able to identify digital objects, physical objects, living beings and groups and intangible objects.[17] The ARK uses a Uniform Resource Locator (URL) scheme to support long-term or permanent access to information objects, and they are sequences of characters following a label“ark:/.”

Digital Object Identifier (DOI)

DOI is a digital identifier of an object, rather than an identifier of a digital object.[18] The scope of the DOIs exceeds the range of digital objects, and they can be used to identify digital, physical and abstract objects. DOI is a typical, domain-independent, identifier system designed by the International DOI Foundation (IDF), which is a non-profit, member-funded organization. IDF has created the DOI system for persistent identification of content with digital environment.[19] DOIs can be assigned to content-related objects, such as text documents, datasets, sound carriers, books, photographs, serials, audio, video, audiovisual recordings, software, abstract works and artwork. An assigned DOI resolves to the bibliographic metadata records of the objects. The metadata records contain current information of the object being assigned the DOI.

Handle System

The Handle System is a domain-independent identifier schema for Internet resources. The Corporation for National Research Initiatives (CNRI) first developed it in 1994, and it was used mainly to resolve DOI. However, Handles also can be used separately. Many institutional repositories use the Handle System as a standalone identification system for research data.[20] Handle System identifiers persist over changes of time, location, ownership and any other conditions.[21] Similar to the scope of DOI, the Handle System is assigned to digital, physical and abstract objects. They resolve to typed metadata records of the assigned objects.

Persistent Uniform Resource Locator (PURL)

PURL was developed by the Online Computer Library Center (OCLC), and it is commonly used as a domain-independent identifier in many institutions.[22] PURL consists of a URL that is a web address that has the feature of persistency. Unlike URLs, which link directly to the locations of Internet resources, PURLs link to middle resolution systems. The PURL Resolution Service maintains the connections between PURLs and their actual URLs and returns the URLs (current locations of resources) to the users. PURLs are linked to metadata records of the assigned objects, such as documents, articles, datasets, web pages and cataloging systems.[23]

Uniform Resource Identifier (URI)

URI is a category of identifier schemas for Web resources. It includes Uniform Resource Locator (URL), and Uniform Resource Name (URN) schemas.[24] URLs specify the location of a resource, URNs specify the name of a resource, which is independent of location. In the classic version (i.e., web of document), the URLs are sufficient as web addresses, although as the locations of web documents change, broken links often occur. However, in the contemporary version (i.e., web of data), which highlights the persistent and unique access of the resource, the condition of non-permanent URLs is no longer sufficient. The changes of the web from classic to contemporary require the use of persistent and unique URIs.[25]

Universally Unique Identifier (UUID)

Originally, UUID was a domain-independent identifier standard used in the computing environment or in software development. The importance of data uniqueness and persistency expanded the usage of UUID from software construction to data identification. UUID supports practical uniqueness guaranteed across space and time.[26] UUID is also generated by its algorithm without a centralized authority, making it less costly. Most other identifiers offer a guaranteed-uniqueness that is administrated via authorities. However, the uniqueness is not unique from a practical perspective if the administrations no longer operate. Conversely, UUIDs are likely to be unique identifiers with their own algorithm, regardless of any authority. Currently, UUIDs are being used within institutional repositories to identify a variety of research data objects and to link to metadata records of the assigned objects.[27]

National Center for Biotechnology Information (NCBI) Accession Number

Since the publication of the human genome project in 2001, biology has entered into a new age within gene and protein sequences.[28] With the advances of the high-throughput sequencing techniques, data on large numbers of genes and proteins must be curated.[29] NCBI's accession number is a unique, domain-dependent, identifier scheme assigned to sequence records when the records are submitted to GenBank, which is a comprehensive database that contains publicly available biological sequence data developed by the NCBI, or to Reference Sequence (RefSeq), which also is a public database for nucleotide and protein sequences synthesized from the sequence data available in GenBank.[30] The accession numbers are unique

numbers that can be embedded in LSID, which is a type of URN, and the embedded number resolves the metadata of the sequence records.

Chemical Abstracts Service (CAS) Registry Number

The number of chemical substances registered in the CAS Registry rapidly increases. According to their report,[31] about 15,000 substances are updated on a daily basis. The CAS Registry contains various types of unique organic and inorganic substances and sequences in their database systems. The substances, such as alloys, coordination compounds, minerals, mixtures, polymers and salts, have distinctive names and structures within the registry. The official titles of substances are used globally to identify the chemical substances. In addition to the CAS Registry, the CAS provides the CAS Registry Number, which is a numeric identifier designed for only one substance.[32] Similar to the NCBI Accession Number, the CAS Registry Number can also be embedded in a URL, and the numbers resolve to metadata records of the chemical substances.

Life Science Identifier (LSID)

LSID is a domain-dependent identifier to identify the entities of life science. Its development was begun in 2003 by the Interoperable Informatics Infrastructure Consortium (I3C). The entities of life science include both concrete and abstract types (e.g., individual proteins or genes, transcripts, experimental datasets, annotations, ontologies, publications and biological knowledge-bases). LSID is an interoperable identifier, so that a namespace, such as an NCBI Accession Number, can be embedded in a LSID, and the LSID can also be embedded in a URN. LSIDs were designed to identify and access biological data in a simple and common way. LSIDs enable their users to access data from various existing resources (e.g., relational databases, applications and public data sources).[33]

International Standard Name Identifier (ISNI)

International Organization for Standardization (ISO) developed ISNI and the specification of valid ISNI standard was published in 2012. ISNI identifies public identities across multiple fields of creative activity. People play in creation, production, management and content distribution chains can be recognized accurately, and the content created from the public identities can be managed effectively. ISNI is

allocated to any party that is or was a natural person, a legal person, a fictional character, or a group or such parties, whether or not incorporated.[34] The assignment of ISNI is based on data aggregated from hundreds of bibliographic and rights management databases, including the Virtual International Authority File (VIAF), which is an international collaborative service to aggregate and provide convenient access to the world's major name authority files.[35] ISNI can also be used as a namespace of URL.

Open Researcher and Contributor ID (ORCID)

In 2012, the ORCID service was launched by the ORCID community and developed to disambiguate scholars with the same name and make connections between research (e.g., research articles and research data) and researchers.[36] The ORCID community maintains it as a registry service, and it has many participants, such as Elsevier and CrossRef.[37] The main goals of ORCID are to provide a reliable identifier and to support its communication and authentication.[38] The format of ORCID is compatible with the format of ISNI.

ResearcherID

ResearcherID was designed by Thomson Reuters in 2008 to solve the ambiguity of authors' names in scholarly communications. Researchers registered with ResearcherID.com are given ResearcherID identifiers. ResearcherID enables researchers to manage their publication lists, check their number of citations, identify future collaborators and avoid author misidentification.[39] Also, ResearcherID information integrates with the data citation index developed by Thomson Reuters, so that researchers can easily discover the publication and its related data from the repository.[40]

OpenID

An open source community trying to solve difficulty of identity metadata managmeent developed OpenID in 2005. OpenID is not limited to the scholarly domain. OpenID is mainly designed for identity authentication for logging on to Web sites. However, it has a potential to be used in open systems as an identifier.[41] People may easily create an OpenID with their preferred OpenID providers. Once they have it, it can be used to facilitate the communication of the users' attributes, such as name and institution, between the provider and the OpenID acceptors.[42]

GeoNameID

GeoNameID is an identifier system used by GeoNames.org. GeoNames is a worldwide database of public geographical data from various sources.[43] It contains more than 10 million geographical names in several layers. In addition, the names of places, latitudes, longitudes, elevations, population and postal codes are stored among its data. The data from GeoNames are freely accessible through various web services.

DATA IDENTIFIER TAXONOMY

Various identifiers exist to support the identification and linking of different types of data in different communities. With the increase of data-driven research and the push for data reuse and sharing by the government, the effectiveness and reuse of metadata schemas, including identifier schemas, gain new importance. Some of the issues, characteristics or contexts related to data identifiers' design, use and evaluation are presented in the following subsections as discussed in the literature.

Domains

Research data can generally be defined as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings."[44] However, the types, formats of and the expertise needed to interpret and curate research data are contextual and domain dependent.[45] In addition, researchers in scientific disciplines more inclined to use domain-specific repositories than institutional general data repositories for their research data.[46]

Various research institutions and communities (e.g., National Center for Biotechnology Information, Chemical Abstracts Service) developed domain-specific identifier schemas to meet their specific needs for identifying and linking datasets, research concepts and entities. At the same time, international or national standard organizations (e.g., International Organization for Standardization, National Information Standards Organization, International DOI Foundation) developed general identifiers that are independent of particular domains.

General identifiers are not limited by disciplines. They have more availability and viability than domain-specific identifiers.[47] Because of their flexible designs, limitations on their uses and assigned entities are lower than other identifiers.

Domain-specific identifiers are designed for particular needs and purposes. To identify the specialized entities of targeted domains, communities analyze data entities and develop their own domain-specific identifiers. Since, these identifiers are tailored to the needs of the domain, they might be less interoperable than general identifier schemas.[48]

Entity Types

Research data may include different types of entities determined by their targeted domains and community norms and policies. Many data repositories store data as application specific computer files. The types of data may include row tabular data, data analysis files, images and drawings, power point presentations and text data files.[49] In addition, community data repositories may also store and maintain knowledge organization tools such as taxonomies, controlled vocabularies, and ontologies, which define different concepts, entities and relationships of the community's knowledge.

A number of researchers[50] have sought to build a map between the identifiers used for traditional library resources (e.g., books, audio-visuals, serials, images) and entity types (in most cases, the FRBR conceptual model's group 1 entities: *work*, *expression*, *manifestation*, and *item*). The map linking identifiers with entity types can be helpful resource in the construction of interoperable data management infrastructure, including data service interoperability, and effective uses of the identifiers.[51]

Several conceptual data models from library, museum, and data preservation communities have been proposed in the literature.[52] The models include entities these communities collect and organize data for. The Open Archival Information System (OAIS) is an ISO conceptual reference model designed to inform the development of systems for long-term digital data curation. The Preservation Metadata: Implementation Strategies (PREMIS) led by the OAIS is a preservation metadata vocabulary. The PREMIS is being widely used in various disciplines.[53] The PREMIS data model consists of five high level entities: intellectual entities, objects, events, agents, and rights.

In the 1990s, libraries faced with a changed information environment that included the variety of data media and new information and data technologies which created new opportunities for more sophisticated uses, aggregation, sharing, analysis and visualization of data in general and bibliographic

data in particular. To support the new uses of bibliographic data, the community needed more systematic model for bibliographic records. The International Federation of Library Associations and Institutions (IFLA) developed and published such a model – the Functional Requirements for Bibliographic Records (FRBR) conceptual model in 1998. The model focuses on supporting four user tasks: to find, identify, select, and obtain bibliographic entities using a library catalog.[54] The FRBR is composed of ten different entities and several relationships among the entities. The ten entities are categorized into three groups. The entities in the first group represent the bibliographic resources in a library catalog. The group entities include work, expression, manifestation, and item. The entities in the second group represent those responsible for the first group's entities. The group's entities contain the person and corporate body. The entities in the third group represent the subject of the first group's entities and include entities of concept, object, event, and place. In the Bibliographic Framework (BIBFRAME) recently developed by the Library of Congress for linked data, bibliographic entities are divided by two classes: creative work and instance.[55]

The International Committee for Documentation (CIDOC) Conceptual Reference Model (CRM) is a formal ontology-supporting museum community developed by the CIDOC of the International Council of Museums (ICOM). The CRM is designed to integrate, mediate, and interchange cultural heritage information.[56] Due to the variety and complexity of information that need to be organized by cultural heritage communities, version 5.1 of the CRM is composed of 90 entities and 152 properties. The following section reviews different data entity types identifiers are used for as referenced in the literature and conceptual models.

Intellectual Entities

The PREMIS define the entity type of Intellectual Entity as "a set of content that is considered a single intellectual unit for purposes of management and description."[57] A book, map, photograph, database, or dataset is the example of the Intellectual Entities. In the FRBR, this type can be mapped to the Group 1 entities (i.e., work, expression, manifestation, and item). To articulate this type of entity, PREMIS used a book *Animal Antics* published in 1902 as an example.[58] A library digitized the book that created one

image file (i.e., TIFF type) for each of 189 pages, and the library also created an XML file to structure the image files. The library also used Optical Character Recognition (OCR) technique on the image files to create a single large text file. The text file was created as an SGML file. The library repository contains *Animal Antics* as an Intellectual Entity that includes two representations, one consisting of 189 image file objects and an XML file object, and the other consisting of one SGML file object (in figure 1). Each representation of the Intellectual Entity is full version of *Animal Antics*.

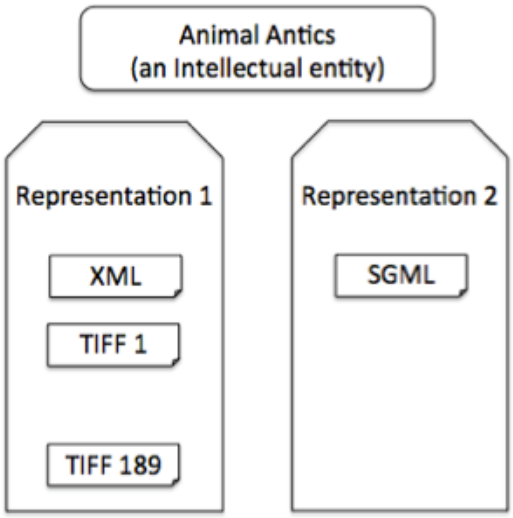

Figure 1. An example of intellectual entity, Animal Antics (LC, 2012).

According to Carlyle,[59] abstract entities such as work and expression make the FRBR difficult to understand for some because their existence is not observable. Discussions on the entities of expression and manifestation have also focused on their ambiguity related to XML documents.[60] Buckland[61] and Floyd and Renear[62] raised the lack of clarity of what is document within the digital environment, reflecting the difficulties of identifying item entity. In addition, Halpin[63] pointed out that identifiers often failed accessing the entity they were assigned to as they directed to the metadata descriptions of the entity, rather than the entity itself. These discussions about the ambiguities and inconsistent use of identifiers with bibliographic entities provide support for the Intellectual Entity being a single entity type for digital content, although it can be mapped to the multiple entities of the FRBR Group 1.[64] PREMIS specified

URI, Library of Congress Control Number (LCCN), and Handle System as the identifier schemas for Intellectual Entities.[65]

Object

Most research data in digital environment exist as computer files or bitstreams. To store the data in digital repositories, the content of the data needs to be digitized. In the PREMIS, the Object is defined as "discrete units of information in digital form."[66] The Object can be thought of as media/carriers of information, such as files, bitstreams, or representations. A dataset (i.e., an example of the Intellectual Entity) can be constructed by many computer files, and the each file is the example of the Object type.[67] As seen in figure 1, each file is an Object. Many different data repositories and data management application tools (e.g., Dryad, DataUp, EZID, etc.) provide platforms for researchers managing and archiving research data. In most cases, the researchers upload their data in targeted repositories via the applications, and they get a unique identifier (e.g., DOI, ARK, etc.) associated with the data.

Symbolic Object

Scientific research data (i.e., an example of Intellectual Entities) in many cases have forms of symbolic representation. Gene and protein sequences in biology and chemical compounds and structures are major examples. Every day scientists discover new DNA strands or chemical substances, and they store the discovery in data repositories and use the data to publish research articles. Alphabetic letters, specialized symbols/signs, etc. describe such scientific objects. The Concept entity in the FRBR is defined as "an abstract notion or idea."[68] Knowledge, theories, practices, and techniques are examples of concept. The Symbolic Object in the CIDOC CRM is an entity type that can be matched with the Concept in the FRBR. The Symbolic Object is defined as identifiable concepts and any aggregation of concepts with an objectively recognizable structure.[69] The examples of Symbolic Object provided by CIDOC CRM include characters, texts, images, computer program codes, mathematical formulae, etc. Accession numbers from GenBank or Reference Sequence (RefSeq) databases assigned to gene or protein sequences can serve as examples of the identifiers used with the Symbolic Object entity type.

Person

For the identification of any digital object, the Person entity is necessary to determine those who create and maintain the objects.[70] Due to the malleable nature of digital data (i.e., easy to modify, aggregate, integrate),[71] metadata about who created, modified and/or accessed a particular data object is essential for discovering the data, and assessing its relevance and quality.[72] The three conceptual models used in this paper include the entity representing human beings. Both the FRBR and CIDOC CRM have the Person entity in the models. In the PREMIS, the Agent entity exists that is defined as actors that affect the information. The Agent can include people, organizations, and software applications. The entities from the three models can help identify research data and control authority data of the assigned research data. As mentioned previously, Elsevier decided to use an author identifier (i.e., ORCID) to create links between scholarly works and their authors. In addition, many social network websites (e.g., Facebook, Twitter, LinkedIn, etc.) provide an alternative identifier option (i.e., OpenID) for their users to integrate public identities and aggregate their content across the Web.

Organization

The Organization is an entity type to identify organizations preserving, managing, or creating research data. The type has similar goals as the person entity, which helps access and retrieve correct research data with controlled authority metadata. This entity type exists in the FRBR and CIDOC CRM. The Corporate Body from the FRBR is the entity corresponding to the Organization. It is defined as an organization or group of individuals.[73] The Legal Body entity corresponds to the organization in the CIDOC CRM. The CIDOC CRM defines the entity as organizations or groups that have obtained legal recognition.[74] In the PREMIS, the organization is embedded in the agent entity. The International Standard Name Identifier (ISNI) designed by ISO is a type of author identifiers, which also identifies organizations as public identities.[75] The Library of Congress uses ISNIs to disambiguate the public parties involved in media content.

Place

Along with the development of the Geographic Information System (GIS), the potential values and uses of geographic data have increased. The data are being actively used in various domains, such as business, economics, history, urban planning, and oceanography.[76] In addition to the GIS data, the importance of the accurate geographic location (i.e., latitude and longitude) as research data has also increased. Knowing the precise location is important to research in oceanology, glaciology, meteorology, etc. The Place entity in the FRBR is defined as a geographical location.[77] The CIDOC CRM's definition of Place is more specific: spatial extents on the surface of the earth.[78] Both models support the geographic location data with this entity. GeoNames is a geographical database, which freely provides over 10 million geographical names and locations to general public. It uses GeoNameID to identify its location data. The identifier schema includes geographical names, latitude and longitude, elevation, timezone, population, etc. as its metadata elements.[79]

Time

It is critical that time information is collected as a part of research data. Recorded time helps find proper and accurate data to meet users' needs. If accurate time records do not exist, researchers might have difficulty identifying and classifying data. For example, if a high-performance camera takes tens or hundreds of photos per second and does not record the exact time each photo was taken, the classification and organization of the data collected by this camera may not be possible. In a real example of the use of camera in space science, researchers organize images of the sun in chronological order to observe and record the rate of changes on its the surface. Such observations have great value as documents and forecasts, and as research data. In the CIDOC CRM, two different entities may convey time information. The entity of Date is defined as specific forms of historical periods or dates, and the entity of Time-Span is defined as abstract temporal extents, having a beginning and an end.[80] Time is essential information in research data, but the literature does not report the use of identifiers with time entities.

Event

With malleable nature of digital, the issue of data reliability and the quality of provenance metadata become even more important. In this context, the importance of all the changes that affect the digital objects is emphasized. The PREMIS defines Event as actions that involve an object and an agent associated with intellectual entities.[81] The Event from the CIDOC CRM effectively reflects the features of the community information (i.e., cultural heritage information) from the definition of the entity. It is defined as changes of states in cultural, social, or physical systems.[82] Finally, the Event from the FRBR is defined as an action or occurrence.[83] The W3C Provenance (PROV) Working Group recently published its model for provenance metadata (i.e., PROV Model). The model defined an Activity entity being compared with Event entity as something that occurs over a period of time and acts upon or with entities. The entity includes the actions of consuming, processing, transforming, modifying, relocating, using, or generating entities.[84] The PROV Model Primer also defines three kinds of provenance perspectives on its users: agent-, object-, and process-centered provenance. The PREMIS supports the provenance of information with the event entity, which focuses on agent- and object-involved information. In research data, the position of provenance information has been particularly emphasized. In a scientific experiment, a small change to an experimental variable can bring about great changes to the experiment's outcome; thus, any change in variables, such as an event, must be accurately recorded. The identifiers designed specifically for Event entity do not currently exist or has not been reported in the literature yet. However, Archival Resource Keys (ARKs) developed by California Digital Library are used to identify many different types of objects and can be used as the identifier for Event entities.

Topic

Topic or subject is an important element of any bibliographic metadata schema. Catalogers, curators, and/or authors assign topic keywords or phrases to resources, which are then used by users to discover relevant data and information resources. Libraries and scholarly communities (e.g., National Center for Biotechnology Information (NCBI), American Institute of Physics (AIP), Library of Congress (LC)) use different thesauri, controlled vocabularies, and ontologies to bring related data together by disambiguating

and reducing vocabulary variance in metadata. For example, in 1970 the AIP developed the Physics and Astronomy Classification Scheme for classifying scientific literature using a hierarchical set of alphanumeric codes. The scheme has been used internationally, including by major physics journals. Google, as it becomes one of the major dataset aggregators, developed the Dataset Publishing Language (DSPL) released the DSPL schema to the public. The DSPL schema requires the use of at least one unique identifier for topic element.[85] The FRBR's third group entities (i.e., Concept, Object, Event, and Place) correspond to the topic element in the DSPL. In the FRBR, the entities in the third group have a bidirectional relationship, entitled "has a subject," with the work entity in the first group.[86] The relationship indicates that the third group entities explain the subjects of creative work. Topic in DSPL schema can be identified and referenced by URIs.

Activities

The study identified four types of activities that use data identifiers: identification, citation, linking, and annotation of research data.

Identification

Identification task can be defined as "confirming that the entity described corresponds to the entity sought [by the user], or distinguishing between two or more entities with similar characteristics."[87] The identification activity is conducted by utilizing identity metadata elements, most importantly identifiers.[88]

Qin, Ball, and Greenberg[89] discussed identity metadata for scientific data. They defined the identity metadata as the properties of entities (e.g., agent, event) that when encoded as metadata can be used to verify the identity of data resources. These entities may also have assigned metadata such as identifiers. For example, author identifiers (e.g., ISNI, ORCID, ResearcherID, etc.) identify agent/person entity, and resource identifiers (e.g., DOI, URI, Handle System, etc.) are assigned to publication, event, and/or dataset entities.

The DataUp project ran by the California Digital Library developed an open-source add-in for Microsoft Excel software. The add-in targeting data management of earth, environmental, and ecological

sciences helps users with documenting and depositing data into a data repository. DataUp add-in uses Archival Resource Keys (ARKs) as a persistent identifier for deposited datasets.[90]

The Organization for Economic Co-operation and Development (OECD) Publishing proposed metadata standard for data publishing.[91] The OECD Publishing specified Digital Object Identifiers (DOIs) as a mandatory identity metadata element for dataset entity. Similarly, Altman and King[92] suggested using unique global identifiers for quantitative research data identification and citation, and they recommended using Uniform Resource Identifiers (URIs) taking Uniform Resource Name (URN) syntax, Life-Science Identifiers (LSIDs), DOIs, and Handle System.

Citation

The main goal of data citation is to build the connection between an identifier and its associated data object at any time in the future,[93] and the minimum component of the connection is a persistent identifier.[94] Many institutional data repositories assign identifiers to data objects to connect them to various types of entities.[95] Citation metadata also can serve as data itself in evaluating the productivity and impact of individual researchers, teams, labs, and communities.[96]

Major funding agencies, such as NSF and NIH, now require applicants to submit plans for managing and providing access to research data.[97] This pressure from funding agencies and their user communities encourages libraries and data centers to establish projects like DataCite to help researchers find, access, and reuse data. DataCite also provides services and tools for data publishers to generate associated metadata. DataCite uses DOIs as its only allowed value of identifiers.[98]

Several other tools/instruments have been developed to help institutions publish, cite, and discover research data. The Dataverse Network (DVN) developed by the Institute for Quantitative Social Science at Harvard University is an open-source application providing useful guidelines and tools for data citation.[99] The application intended to motivate researchers to share data through enabling persistent data citation using a global persistent identifier with URL, and universal numerical fingerprint. The DVN specifies Handle System and its Global Handle Registry as their persistent identifier system. Also, DOIs,

which use Handle System infrastructure on their name resolution, can easily be used as the standard identifier system with the DVN application.

The Data Observation Network for Earth (DataONE) is a National Science Foundation (NSF) supported project, which intends to improve access to, and preserve data. The DataONE community developed a method for data citation on the areas of life and earth science. The Dryad repository – a member repository of the DataOne - asks its users who cite data in Dryad to use either DOIs or ARKs. DOIs used by the Dryad are registered at DataCite, and the DOI registration information contain data citation metadata elements required by the Dryad (i.e., author(s), date, title of the data package, repository name, and data identifier).[100]

Linking

The activity of linking can be defined as the connection between data that was not previously linked, or the connection of data lowering the barriers to linking data currently linked using other methods.[101] W3C introduced the concept of linked data in 2006. It is defined as a set of best practices for publishing and connecting data on the web.[102] In brief, data is serialized and published on the Web using the Resource Description Framework (RDF) based format, which potentially allows to connect the data with other related datasets at a low cost. Linked data are not just about uploading data on the web, but about generating links.[103]

Linked data principles outlined by Berners-Lee[104] emphasize the use of HTTP URI. A datum is represented by a URI, and the two related URIs are linked by another URI. The three URIs accordingly form a RDF triple.[105] If identifiers are used as HTTP URIs (in table 1), it is possible to generate RDF links.[106]

If research data in a data repository are not associated with relevant articles, the data is hidden, limiting its use and reuse. The frequency of data use can be closely related to the value of the data, and the value can be improved by connecting them to entities of relevant articles.[107] The entities, in this context, can be defined as discipline-specific concepts used in the research.

Elsevier currently provides linking services that aim to add value to scientific articles. Datasets are connected to articles (i.e., dataset linking) or to the entities of articles (i.e., entity linking) by identifiers. The dataset linking service makes linking based on the DOI assigned to an article. The entity linking service accepts various accession numbers from various databases (e.g., GenBank, Protein Data Bank, Cambridge Crystallographic Data Centre, Molecular Interactions Database, and Universal Protein Resource Knowledgebase) with URI syntax as its identifiers.[108] Elsevier also stores data as RDF documents in Linked Data Repository.

Table 1. Data identifiers as URIs with URL/URN syntax.

Identifiers	URI		Sources
	URL	URN	
ARK	Yes		CDL, 2012
DOI	Yes	Yes	DOI, 2013
Handle System	Yes		CNRI, 2012
PURL	Yes		OCLC, n.d.
UUID		Yes	Leach et al., 2005
NCBI Accession Number		Yes, with LSID	Clark et al., 2004
CAS Registry Number	Yes		Common Chemistry, 2013
LSID		Yes	Clark et al., 2004
ISNI	Yes		ISNI, 2012
ORCID	Yes		ORCID, n.d.
ResearcherID	Yes		ResearcherID, n.d.
OpenID	Yes		OpenID Foundation, 2007
GeoNameID	Yes		Pabón et al., 2013

Annotation

Annotation is a process of adding notes on or commentary to informational sources. Annotations may enhance the value of data by connecting or supplementing it with relevant descriptions, explanations and interpretations.[109] Annotating and integrating research data with relevant scholarly works tend to rapidly increase with data-driven research in scientific disciplines.[110]

The National Center for Biotechnology Information (NCBI) developed Reference Sequence (RefSeq) database, which has authority over biological sequences within the GenBank database. The

RefSeq is used by biological scientists as an authority file by having access to well annotated genomic DNA, transcripts, and protein sequences.[111] RefSeq uses its accession number as identifiers for scientific annotations.

W3C Open Annotation Community Group recently published Open Annotation Data Model, which provides a framework for annotation. The framework proposes the open annotation following the linked data principles.[112] The annotation is considered to be a set of linked resources including "body" and "target." In most cases, the body explains the target. The model recommends to use URIs to make connections between the resources.

Quality Dimensions

To support the activities of identifiers and evaluate their quality, many researchers have suggested or developed different quality requirements.[113] Quality is usually defined as "fitness for use."[114] Quality is multidimensional and contextual and there could be tradeoffs among different quality dimensions.[115] Table 2 shows the definitions of the quality dimensions and sources referencing the dimensions. In the following, we will discuss seven quality dimensions in more detail: simplicity, opacity, verifiability, contextuality, interoperability, actionability, and granularity.

Table 2. Description of the quality dimensions and their sources.

Dimensions	Definitions	Sources
Uniqueness	The requirement that one identifier string denotes one and only one data object	Altman & King, 2007; Michener et al., 2011; Lagoze et al., 2006; Paskin, 2010; Weigel et al., 2013
Persistence/ Volatility/ Legacy support	The requirement that once assigned, an identifier string denotes the same referent indefinitely	Altman & King, 2007; Berners-Lee, 1998; Brand et al., 2003; Callaghan et al., 2012; Duerr et al., 2011; Michener et al., 2011; NISO/NFAIS, 2013; Lagoze et al., 2006; Paskin, 2010; Tonkin, 2008; Vitiello, 2004; Weigel et al., 2013
Simplicity/Transparency	The degree of cognitive simplicity of an identifier string	Berners-Lee, 1998; Duerr et al., 2011; NISO/NFAIS, 2013; Tonkin, 2008
Opacity	The extent to which the meaning can be inferred from the content, structure or pattern of an identifier string	Brand et al., 2003; Clark, 2006; Duerr et al., 2011; Michener et al., 2011; NISO/NFAIS, 2013; Tonkin, 2008
Verifiability	The extent to which the correctness and validity of an identifier string is verifiable or provable	Akhondi el al., 2012; Duerr et al., 2011; Juty et al., 2012; Tonkin, 2008

Contextuality	The degree of an identifier system and string for the needs of a targeted community	Clark et al., 2004; Juty et al., 2012; Tonkin, 2008
Compatibility	The ability to use with the main internet naming schemes (i.e., URL or URN)	Duerr et al., 2011
Interoperability	The ability to use an identifier system and string in services outside of the direct control of the issuing assigner	Altman & King, 2007; Berners-Lee, 1998; Duerr et al., 2011; NISO/NFAIS, 2013; Paskin, 2010; Vitiello, 2004
Actionability /Resolvability	The ability of the identifier system to locate the object using an identifier string	Altman & King, 2007; Brand et al., 2003; Callaghan et al., 2012; Duerr et al., 2011; Juty et al., 2012; Michener et al., 2011; NISO/NFAIS, 2013; Lagoze et al., 2006; Paskin, 2010; Tonkin, 2008; Vitiello, 2004; Weigel et al., 2013
Granularity/ Flexibility	The extent to which the identifier system allows to reference data at different granularity	Juty et al., 2012; Michener et al., 2011; Tonkin, 2008; Vitiello, 2004;
Authority	The degree of reputation of an identifier system in a given community	Altman & King, 2007; Duerr et al., 2011; NISO/NFAIS, 2013; Tonkin, 2008
Scalability	The ability of an identifier system to expand its level of performance or efficiency (e.g., support RDF)	Duerr et al., 2011; Juty et al., 2012; Lagoze et al., 2006
Security	The extent to which the resource of an identifier system is protected from unauthorized administrative access or modification	Duerr et al., 2011; Juty et al., 2012; Tonkin, 2008

Simplicity/Transparency & Opacity

Identifiers within different contexts have different requirements on their strings. In the context of data aggregation, communities prefer transparent and simple strings.[116] Information about the characteristics of data objects encoded in identifier strings in a transparent way can be helpful in the disambiguation, aggregation, or clustering of data objects along those characteristics. On the other hand, when data is sensitive, opaque identifier strings are preferred.[117] Opaque identifiers could be more robust as they are not sensitive to changes in the characteristics of data (e.g., entity name change).[118]

Verifiability

Identifier strings often have a complex syntax. The complexity causes various issues related to verification and validity of the strings. Often checksums or other error-correction mechanisms are used to ensure identifier string validity. Identifier string verification for digital resources can be relatively simpler

than the one for physical resources. Network connection might provide a quick solution, checking the correctness or validity of the strings by returning the associated data objects.

Contextuality

Many identifier systems are developed to meet specific community's needs. Data-driven research trends also accelerate the use and development of community-driven identifiers and repositories.[119] The large amount of and various types of research data require more sophisticated curation, including the development of identifiers schemas which are tailored towards the community's data management needs.[120] In addition, an identity tension exists on determining the type (i.e., domain and entity type) of identifiers. For example, the domain of a URI related to Caffeine might be chemistry, pharmacy or nutrition. Berners-Lee[121] takes a position that the type of a URI is whatever the owner intended. Halpin[122] has emphasized the importance of communities' names, combination of words and signs as a shared mode of presentation to define the type of a URI. Understanding of community context/use is essential for determining the domain of an URI. Also, the entity type of the URI can be Symbolic Object or Topic based on its community use. Hayes[123] takes similar, but slightly different approach arguing that the type of a URI is determined by linked structured resources (i.e., RDF triples) within Semantic Web.

Interoperability

Interoperability aiming a shared understanding of data can be defined as the exchange and use of information in an efficient and uniform manner across multiple organizations and systems.[124] In this context, Paskin[125] identified three distinguished identifier interoperability in the aspects of syntax, semantics, and community. Syntactic interoperability is the ability of systems to read and recognize more than one identifier syntax string within an identifier string. For example, LSIDs use a form of URN and can include an identifier string, such as NCBI Accession Number, within their syntax strings. Semantic interoperability is the ability of systems to determine how two associated data objects are semantically related. It can be conducted by using structured metadata or ontologies. For example, ontologies and conceptual models such as the CIDOC CRM, Online Information Exchange (ONIX), and Resource Description and Access (RDA) can be used in semantic integration.[126] Finally, the community

interoperability is the ability of systems to collaborate and communicate between different identifier systems without any restrictions of each system use. The community interoperability first requires community policies that are willing to share and compare their metadata management plans with other communities. Otherwise, the interoperability can not be viable. According to Paskin, these three aspects are dependent. Syntactic interoperability is a required condition of ensured semantic interoperability, which is necessary to ensure community interoperability. Pabón et al.[127] mentioned that all of "legal compatibility, semantic interoperability, technical aspects of information systems, organisational cooperation and a favourable political climate" are necessary for interoperable services in reality.

Actionability/Resolvability

In general, resolution systems are a bridge system including both input and output. The input is an identifier string as a key, and the output is the current information associated with the identified objects.[128] It is strongly recommended identifier systems to have a resolver to track down dynamic locations of data objects. An identifier system with a resolver does not require any change of identifier strings, even when the physical location of the identified object is changed.

Granularity

Research can be driven by multiple research data. A dataset usually contains multiple data files. According to Lee and Stvilia,[129] institutional repositories store multiple types of research data files (e.g., single data files, compressed data files, and database files). The need for multiple granularity identification happens when a researcher wants to cite only one specific file from a dataset.[130] For example, if a dataset contains one hundred files and the researcher wants to cite only one file in that dataset. To support this need, the identifier system needs to support data referencing at multiple granularity.

PRACTICAL USE OF DATA IDENTIFIERS

The taxonomy from Figure 2 presents a summary of the concepts related to identifier schema design, use and evaluation as discussed in the above sections. The taxonomy consists of four main categories and their sub-elements.

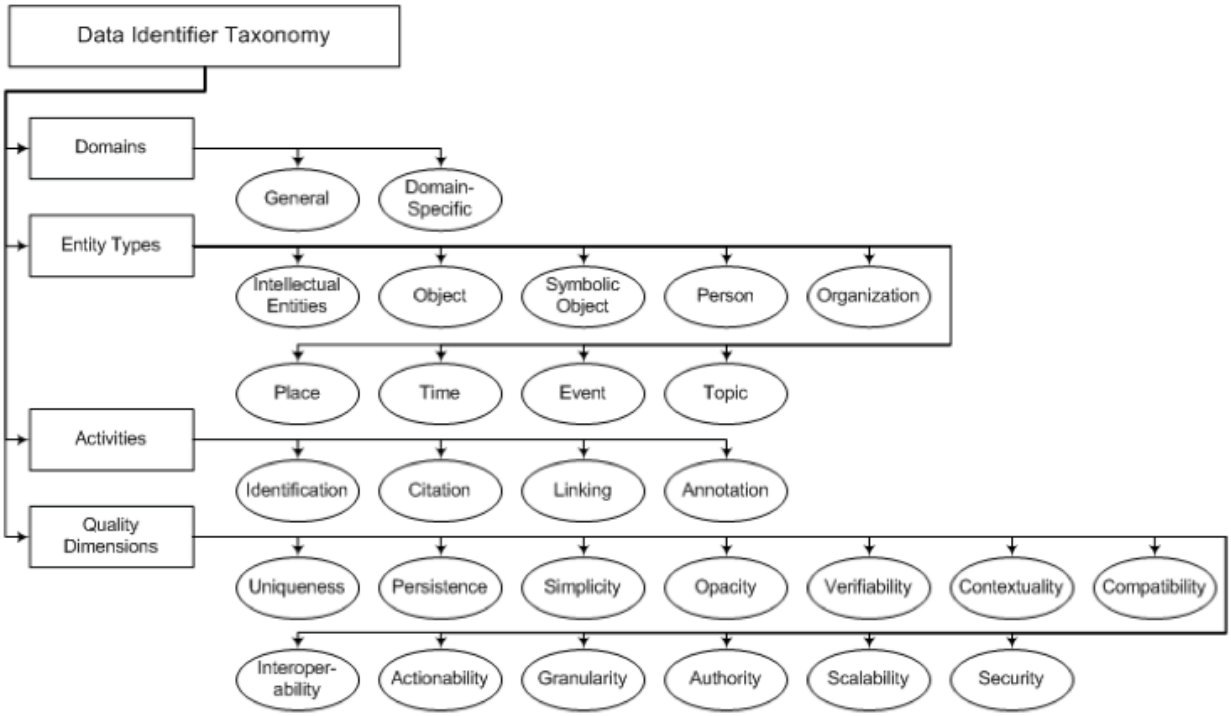

Figure 2. Data identifier taxonomy.

In a next step, the study examined the practice of current data identifiers based on the characteristics defined by the taxonomy. The 14 identifiers reviewed in the previous section were selected for an analysis, and the conceptual analysis of those identifiers was conducted based on technical specifications, user documentation, and published journal articles. Table 3 briefly summarizes the results of the analysis. The empty cells within the table indicate the absence of a particular property or use, and the cells marked with "Yes" indicate the opposite. This analysis has limitations. In some cases, the literature used in this analysis provided clues rather than a direct answer for individual cells, and many results obtained from the literature do not include comparative elements among the identifiers. The results provide a conceptual understanding based on the literature analysis and require further research using empirical data. In this section, we discuss the results of this analysis.

Domains and Entity Types

Six identifiers were identified as domain-independent identifiers: ARKs, DOIs, Handles, PURLs, URIs, and UUIDs. These are primarily assigned to the Intellectual Entities within many institutional data

repositories,[131] and some of them (i.e., ARK and URI) can be assigned to author- and subject-related entities.

Before the mapping of the identifiers to Intellectual Entity, it is worth reasoning about the distinctions and issues between individual FRBR Group 1 entities. The distinctions are not unambiguously defined, as we previously mentioned for the abstract entities in the digital environment.[132] This conceptual ambiguity can be even worse with research data. Hence, mapping the domain-independent identifiers to the broader Intellectual Entity (i.e., a broader entity, including abstract entities) rather than individual FRBR Group 1 entities can be more meaningful and robust. Previous similar studies have not mapped the identifiers to the individual FRBR Group 1 entities, or concluded that the mapping is meaningless.[133]

Table 3. Summary of practices of data identifiers.

	ARK	DOI	Handle System	PURL	URI	UUID	NCBI Accession Number	CAS Registry Number	LSID	ISNI	ORCID	Research-erID	Open ID	GeoName ID
Domains														
General	Yes	Yes	Yes	Yes	Yes	Yes				N/A	N/A	N/A	N/A	N/A
Domain-Specific							Yes	Yes	Yes	N/A	N/A	N/A	N/A	N/A
Entity Types														
Intellectual Entities	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes					
Object	Yes	Yes	Yes	Yes	Yes	Yes			Yes					
Symbolic Object	Yes				Yes		Yes	Yes	Yes					
Person	Yes				Yes					Yes	Yes	Yes	Yes	
Organization	Yes				Yes					Yes				
Place	Yes				Yes									Yes
Time														
Event	Yes				Yes									
Topic	Yes				Yes									
Activities														
Identification	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Citation	Yes	Yes	Yes	Possibly	Possibly	Possibly	Possibly	Possibly	Yes	Possibly	Possibly	Possibly	Possibly	Possibly
Linking	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Annotation	Possibly	Yes	Possibly	Possibly	Yes	Possibly	Yes	Yes	Yes	Possibly	Possibly	Possibly	Possibly	Yes
Quality Dimensions														
Uniqueness	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Persistence	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Simplicity	Yes	Some	Some	Yes	Yes				Yes				Yes	
Opacity	Yes	Yes	Yes	Yes	Yes	Very	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Verifiability	Very	Yes	Yes	Yes	Yes		Yes	Very	Yes	Very	Yes	Yes	Yes	Yes
Contextuality							Yes	Yes	Yes					
Compatibility	Yes	Yes	Yes	Yes	N/A	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Interoperability		Yes	Yes						Yes		Yes			
Actionability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Granularity		Yes												
Authority	Some	Very	Very	Some	Very	Some	Very	Very	Very	Very	Very	Very	Some	Some
Scalability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Security	Very	Very	Very	Some	Some	Very	Very	Very	Very	Very	Very	Very	Some	

ARKs and URIs can also be used with entities in other groups—namely, author and subject. ARKs can be assigned to various types of objects (e.g., digital, physical, and intangible objects and living beings and groups) with flexible and wide range of scopes. URIs are compatible with all the identifiers (in table 1).

In biology, alphabetic letters express gene or protein sequences. Accession numbers from the NCBI assigned to the expressed alphabetic records can be mapped to Intellectual Entities and Symbolic Objects. The sequence records can be considered as intellectual concepts or the symbolic expression of intellectual concepts.

CAS Registry Numbers are associated with molecules of chemical substances, which are the smallest amount of a chemical substance. In most cases, the molecules are intangible and invisible to the naked human senses. An object assigned a CAS Registry Number is, therefore, a molecular expression of the substance written by chemical formulas and symbols. Similar to NCBI Accession Numbers, CAS Registry Numbers can be matched with Intellectual Entity and Symbolic Object.

LSID is an identifier associated with data resources related to life sciences. The data include both concrete and abstract objects. LSID has wide scopes similar to domain-independent identifiers, but it is only applied to the resources in life sciences. It can be associated with protein or gene sequences by cooperating with various namespaces (e.g., GenBank, Protein Data Bank (PDB), GeneOntology) and data files in the field of life sciences.[134] LSIDs can be associated with Intellectual Entity, Object, and Symbolic Object.

The identifiers (i.e., ISNI, ORCID, ResearcherID, and OpenID) designed to associate with people or groups can be mapped to Person or/and Organization entities. Although they work in similar ways, some differences exist in their uses. ISNIs can be linked to the public identities of any kind of producers of intellectual content, including the names of organizations. ORCIDs and ResearcherIDs can only be associated with researchers and research contributors. Lastly, OpenIDs—used mostly for the purpose of identity authentication—can be assigned only to persons.

GeoNameID is an identifier that identifies accurate geographic location.

Activities

The researchers examined the use of identifier systems with data and scholarly activities (in table 3). If the use of the identifier system in a particular activity system was mentioned in the literature, the corresponding cell was marked with "Yes" in table 3. "Possibly" means that the identifiers seem to be applicable, but their use has not been reported in the literature. All the identifiers fully support the activities of identification. The linking activity too can be supported by all the identifiers. All of them can be used as a URI (one of the requirements of linked data implementation) (in table 1). ARKs, DOIs, Handle Systems, and LSIDs are currently used as identifiers in different data citation models. The rest of the identifiers with URL- or URN-syntax too can be used as the identifiers in data citation[135] but their use has not been reported in the literature yet. For example, the American Psychological Association's (APA) publication manual only allows DOI and URL as citable identifiers.[136] This policy, however, is quite flexible and means that any identifier, which has the URL syntax, can be used in data citation. Three domain specific identifiers (i.e., NCBI Accession Numbers, CAS Registry Numbers, and LSIDs), DOIs, URIs, and GeoNameIDs are currently used within annotation activity.[137] The other identifiers do not seem to have any barriers for supporting annotation, but their use within the annotation activity has not been reported in the literature.

Quality Dimensions

The study analyzed the designs and specifications of the fourteen identifier schemas to assess their quality along the seven quality dimensions. A simple three level scale ("Very," "Yes," and "Some") was used to indicate their relative rankings along those dimensions. The ARK, PURL, URI, LSID, and OpenID allow their schema users to generate the identifier strings according to their own rules—a flexibility that enables their users to tradeoff between two conflicting quality dimensions: transparency and opacity. For instance, if a string meet with the minimum requirements to be the schema string, the remaining part of the string can be created by user to meet local data management priorities and preferences. DOI and Handle System too permit their schema users to create a part of an identifier string—namely, the suffix of the string. Hence DOI and Handle system strings can be made transparent or simple.

All the identifier schemas except UUID are verifiable with an Internet connection. In addition, the verifiability of ARK, CAS Registry Number, and ISNI identifiers is supported by checksum functions. The quality of interoperability is important for the purpose of identifier synthesis.[138] DOI is interoperable with the Handle system and many ISO identifiers (e.g., ISBN, ISSN, etc.). The Handle system shares much of the technology (e.g., protocol) with DOI.[139] URI is compatible with the identifier schemas which have been used with URI schemas (i.e., URL and/or URN) (in table 1). LSID includes a name authority within its syntax, such as the NCBI Accession Number embedded in LSID. ORCID shares its syntax with ISNI. The granularity is one of the difficult quality dimensions. Most identifiers do not fully support multiple granularities. However, DOI has been used to support data identification and access at multiple granularities at Dryad, which is a repository for research data in biosciences.[140] At Dryad, suffixes of assigned DOIs are generated by their own rules, displaying the relationship between data collection and a single data file within the collection. Similar to the Dryad DOI profile, the other identifier schemas that rank high on the simplicity dimension can be used to support data identification and access at multiple granularities.

DISCUSSION

This study used a literature review to develop a taxonomy of identifier schemas used for research data. The taxonomy consists of four categories: domains, entity types, activities, and quality dimensions (in figure 2). It can help data curators in the selection of appropriate identifier systems for their data repositories and the development of data management and citation policies. Although many researchers have studied the issues of design and the use of identifier schemas, to the best of our knowledge Duerr et al.'s[141] study was the only one previous study to include a comprehensive and systematic review of the current identifier systems.. Duerr et al. examined the utility of identifier schemas for digital earth science data. The assessment was conducted with 14 assessment criteria categorized as technical value, user value, or archive value. This study synthesized an identifier taxonomy based in a literature review which included Duerr et al' work. Hence, our taxonomy included all of the quality criteria from Duerr et al. However, it takes an activity perspective on quality. In addition to compiling a list of identifier quality

dimensions mentioned in the literature, the taxonomy makes a focus on the relationships among activities, entities, and identifier quality dimensions. The value and usefulness (i.e., quality) of an identifier schema is ultimately defined by its ability to meet the activity's requirements. For example, our taxonomy includes citation as an activity. Data citation activity then requires having an identifier system satisfying specific quality characteristics (e.g., uniqueness, actionability, persistence, interoperability)[142] (in table 2). Duerr and colleagues used publisher's acceptance of an identifier in a citation as a quality criterion and grouped it under the user value category. The other criteria (actionability, authority, security, scalability, interoperability, compatibility, persistence, verifiability, simplicity, and opacity) from the Duerr et al. study are included in the list of quality dimensions of our taxonomy. The taxonomy also includes quality dimensions which are not a part of Duerr et al.'s values assessment model (e.g., contextuality and granularity).

The taxonomy includes a typology of activities: the activities of identification, citation, linking, and annotation for data identifiers. The first two can be compared with the use cases of unique identifier and citable locator from Duerr et al. The study dealing with different domains and different data entities required extra activities-related connections among data or between data and original articles through various disciplines. The study for the relevant data identifier design and use in different domains, compared to the study investigating utility of identification schemes in a specific domain, provided a clear distinction between activities and quality dimensions and additional activities related to data integration.

The unique contributions of this study include the identification of data entity types and their mapping to data identifiers. Although many researchers have constructed a map between the identifiers used for traditional library resources and the entity types of the FRBR conceptual model, they have not conducted the mapping on research data entities with identifiers. Vitiello[143] and LeBoeuf[144] reviewed various identification systems (i.e., ISTC for textual works, ISWC for musical works, ISAN for audiovisual resources, ISRC for sound and music video recordings, ISBN for books, ISMN for music publications, ISSN for serials, SICI for serials items, ISRN for reports, and DOI for digital object) and mapped them to the first group of entities of the FRBR conceptual model. Another study presented at the

European Library Automation Group's 2010 conference extended LeBoeuf's work with additional mappings of identifiers (i.e., OWI for OCLC works, ISCI for collections, ISNI for names, ORCID for researchers, OCN for OCLC records, NBN for bibliographic resources in the National Library of Finland, and LCCN for resources in the Library of Congress) and entities (i.e., authors).[145] This study analyzed all these valuable previous work, the identifiers and entity types they discussed, and reflected them in its literature review and taxonomy (i.e., ISNI, ORCID, DOI, work, expression, and manifestation as intellectual entities and authors). At the same time, this study extended the scope of the previous work to research data and developed a more comprehensive typology of entity types (in figure 2).

The analysis of current practices of identifier use for research data point to several issues (in table 3). Current identifier systems are not sufficient for identifying all of the data concepts. In particular, the literature suggested that current identifier schemas could provide only limited support the entities, activities, and quality dimensions related to data provenance[146] (i.e., agent, place, time, event, topic, annotation, granularity, simplicity, and scalability.) Time entity had not been supported by any identification schemas. Annotation activity seemed to need the development of policies and practices for data representation in multiple domains. In addition, the need for multiple granularity data identification and access remained the most challenging quality requirement for the identifier schemas (in table 3) Finally, the balkanization of research data and ensuring of the interoperability of identifier schemas used with research data remained as significant challenge.[147]

CONCLUSION

Using a literature review this paper developed a taxonomy which can be used as a knowledge source for understanding and reasoning about data identifiers currently used in different domains for different data entities and activities. The study also identified several issues and open research questions. In particular future studies could develop activity specific quality models for data identifier systems in different domain, including developing metrics for the quality dimensions identified in this study. The taxonomy can also guide the design of new data identifier schemas.

The study findings can also inform librarians, repository managers, data curators, scholarly communities, system developers, and publishers about the needs and requirements for an identifier schema to help identify, cite, link, and annotate research data as well as the issues and problems related to the current uses of identifiers for data.

NOTES

[1] National Institutes of Health, "NIH Data Sharing Policy and Implementation Guidance (NIH Publication No. 03-05-2003)," 2010, http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm; National Science Foundation, "Grant Proposal Guide (gpg 11001)," 2010, http://www.nsf.gov/publications/pub_summ.jsp?ods_key=gpg.

[2] Ijsbrand Jan Aalbersberg and Ove Kähler, "Supporting Science through the Interoperability of Data and Articles," *D-Lib Magazine* 17, no. 1/2 (January 2011), doi:10.1045/january2011-aalbersberg.

[3] Jan Brase and Adam Farquhar, "Access to Research Data," *D-Lib Magazine* 17, no. 1/2 (January 2011), doi:10.1045/january2011-brase.

[4] National Science Foundation, "Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans (NSF 10-077)," May 2010, http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928; Office of Science and Technology Policy, "Expanding Public Access to the Results of Federally Funded Research | The White House," 2013, http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research.

[5] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon, "A Survey of Data Provenance in E-science," *SIGMOD Record* 34, no. 3 (2005): 31–36, doi:10.1145/1084805.1084812.

[6] Micah Altman and Gary King, "A Proposed Standard for the Scholarly Citation of Quantitative Data," *D-Lib Magazine* 13, no. 3/4 (April 2007).

[7] Ruth Duerr et al., "On the Utility of Identification Schemes for Digital Earth Science Data: An Assessment and Recommendations," *Earth Science Informatics* 4, no. 3 (2011): 139–160, doi:10.1007/s12145-011-0083-6; NISO, *Improving OpenURLs Through Analytics (IOTA): Recommendations for Link Resolver Providers*, 2013, http://www.niso.org/apps/group_public/download.php/10811/RP-21-2013_IOTA.pdf.

[8] Besiki Stvilia et al., "Research Project Tasks, Data, and Perceptions of Data Quality in a Condensed Matter Physics Community," *Journal of the American Society for Information Science and Technology* (in press).

[9] Shuheng Wu, Besiki Stvilia, and Dong Joon Lee, "Authority Control for Scientific Data: The Case of Molecular Biology," *Journal of Library Metadata* 12, no. 2–3 (2012): 61–82, doi:10.1080/19386389.2012.699822.

[10] Saber A. Akhondi, Jan A. Kors, and Sorel Muresan, "Consistency of Systematic Chemical Identifiers Within and Between Small-molecule Databases," *Journal of Cheminformatics* 4, no. 1 (December 13, 2012): 35, doi:10.1186/1758-2946-4-35.

[11] Thomson Reuters, "Thomson Reuters Unveils Data Citation Index for Discovering Global Data Sets," June 2012, http://thomsonreuters.com/content/press_room/science/686112.

[12] Angela Guess, "Elsevier Joins ORCID in Launch of ORCID Registry," *Semanticweb.com*, October 2012, http://semanticweb.com/elsevier-joins-orcid-in-launch-of-orcid-registry_b32762.

[13] Altman and King, "A Proposed Standard for the Scholarly Citation of Quantitative Data."

[14] Sam Pepler and Kevin O'Neil, *Preservation Intent and Collection Identifiers*, 2008, http://epubs.cclrc.ac.uk/bitstream/2359/Report_II_PreservationIntentAndCompoundObjectIdentifiers-1.pdf.

[15] NISO/NFAIS, *Recommended Practices for Online Supplemental Journal Article Materials*, 2013, http://www.niso.org/workrooms/supplemental.

[16] Dong Joon Lee and Besiki Stvilia, "Identifier Schemas and Research Data," *Proceedings of the American Society for Information Science and Technology* 49, no. 1 (2012): 1–4, doi:10.1002/meet.14504901311.

[17] California Digital Library (CDL), "ARK (Archival Resource Key) Identifiers," 2012, https://wiki.ucop.edu/display/Curation/ARK.

[18] Norman Paskin, "Digital Object Identifier (DOI®) System," in *Encyclopedia of Library and Information Sciences, Third Edition* (Taylor & Francis, 2010), 1586–1592, http://www.tandfonline.com/doi/abs/10.1081/E-ELIS3-120044418.

[19] Ibid.

[20] Lee and Stvilia, "Identifier Schemas and Research Data."

[21] Corporation for National Research Initiatives (CNRI), "System Fundamentals," October 2012, http://www.handle.net/overviews/system_fundamentals.html.

[22] Keith Shafer et al., "Introduction to Persistent Uniform Resource Locators," n.d., http://purl.oclc.org/docs/long_intro.html.

[23] Ibid.

[24] World Wide Web Consortium (W3C), "URIs, URLs, and URNs: Clarifications and Recommendations 1.0," 2001, http://www.w3.org/TR/uri-clarification/.

[25] Tim Berners-Lee, "Cool URIs Don't Change" (W3C, 1998), http://www.w3.org/Provider/Style/URI.html.

[26] Paul Leach, Michael Mealling, and Rich Salz, "A Universally Unique IDentifier (UUID) URN Namespace" (Internet Engineering Task Force (IETF), 2005), http://www.ietf.org/rfc/rfc4122.txt.

[27] Lee and Stvilia, "Identifier Schemas and Research Data."

[28] Paul Higgs and Teresa Attwood, *Bioinformatics and Molecular Evolution* (Malden, MA: Blackwell Publishing Company, 2005).

[29] W. John MacMullen and Sheila O. Denn, "Information Problems in Molecular Biology and Bioinformatics," *Journal of the American Society for Information Science and Technology* 56, no. 5 (2005): 447–456, doi:10.1002/asi.20134; Wu, Stvilia, and Lee, "Authority Control for Scientific Data."

[30] Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott, "NCBI Reference Sequence (RefSeq): a Curated Non-redundant Sequence Database of Genomes, Transcripts and Proteins," *Nucleic Acids Research* 33, no. Database Issue (January 1, 2005): D501–D504, doi:10.1093/nar/gki025.

[31] Chemical Abstracts Service (CAS), "CAS Registry and CAS Registry Number FAQs," 2012, http://www.cas.org/content/chemical-substances/faqs.

[32] Chemical Abstracts Service (CAS), "CAS Registry - The Gold Standard for Chemical Substance Information," 2012, http://www.cas.org/content/chemical-substances.

[33] Tim Clark, Sean Martin, and Ted Liefeld, "Globally Distributed Object Identification for Biological Knowledgebases," *Briefings in Bioinformatics* 5, no. 1 (March 2004): 59–70.

[34] ISNI, "ISNI," 2012, http://www.isni.org/.

[35] OCLC, "VIAF," accessed December 10, 2013, http://www.oclc.org/viaf.en.html.

[36] ORCID, "What Is ORCID?," n.d., http://about.orcid.org/about/what-is-orcid.

[37] CrossRef, "CrossRef & ORCID," 2011, http://www.crossref.org/01company/orcid.html.

[38] ORCID, "What Is ORCID?".

[39] ResearcherID, "What Is researcherID?," n.d., http://www.researcherid.com/Home.action?returnCode=ROUTER.Unauthorized&SrcApp=CR&Init=Yes.

[40] Thomson Reuters, "Thomson Reuters Unveils Data Citation Index for Discovering Global Data Sets."

[41] Simeon Warner, "Author Identifiers in Scholarly Repositories," *arXiv:1003.1345* (March 5, 2010), http://arxiv.org/abs/1003.1345.

[42] OpenID Foundation, "OpenID," 2013, http://openid.net/.

[43] GeoNames, "About GeoNames," n.d., http://www.geonames.org/about.html.

[44] Office of Management and Budget, "Uniform Administrative Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations (OMB Circular 110)" (The White House, 1999), http://www.whitehouse.gov/omb/circulars_a110#36.

[45] Hong Huang et al., "Prioritization of Data Quality Dimensions and Skills Requirements in Genome Annotation Work," *Journal of the American Society for Information Science and Technology* 63, no. 1 (2012): 195–207, doi:10.1002/asi.21652.

[46] Ricky Erway, *Lasting Impact: Sustainability of Diciplinary Repositories* (Dublin, Ohio: OCLC Research, 2012), http://www.oclc.org/research/publications/library/2012/2012-03r.html.

[47] Emma Tonkin, "Persistent Identifiers: Considering the Options," *ARIADNE* 56 (2008), http://www.ariadne.ac.uk/issue56/tonkin.

[48] Wu, Stvilia, and Lee, "Authority Control for Scientific Data."

[49] Besiki Stvilia et al., "Studying the Data Practices of a Scientific Community," in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13 (New York, NY, USA: ACM, 2013), 425–426, doi:10.1145/2467696.2467781.

[50] European Library Automation Group (ELAG), "Workshop on FRBR and Identifiers" (presented at the European Library Automation Group 2010, Helsinki, Finland, June 2010), http://elag2010.nationallibrary.fi/files/2010/06/ELAG-2010-workshop-on-FRBR-and-identifiers.pdf; Patrick LeBoeuf, "Identifying 'Textual Works': ISTC: Controversy and Potential" (presented at the FRBR in 21st Century Catalogues: An Invitational Workshop, Dublin, Ohio, May 2005), http://www.oclc.org/research/activities/frbr/frbr-workshop/program.html.

[51] Thomas Baker and Makx Dekkers, "Identifying Metadata Elements with URIs," *D-Lib Magazine* 9, no. 7/8 (July 2003), doi:10.1045/july2003-baker; Laura Wynholds, "Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects," *International Journal of Digital Curation* 6, no. 1 (March 11, 2011), doi:10.2218/ijdc.v6i1.183.

[52] Library of Congress, "PREMIS Data Dictionary for Preservation Metadata, Version 2.2.," 2012, http://www.loc.gov/standards/premis/v2/premis-2-2.pdf; ICOM/CIDOC CRM SIG, *Definition of the CIDOC Conceptual Reference Model*, 2012, http://www.cidoc-crm.org/docs/cidoc_crm_version_5.1.pdf; International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records*, 2009, http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

[53] Library of Congress, "Premis Implementation Registry," 2011, http://www.loc.gov/standards/premis/registry/.

[54] International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records*.

[55] Library of Congress, "Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services," 2012, http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf.

[56] ICOM/CIDOC CRM SIG, *Definition of the CIDOC Conceptual Reference Model*.

[57] Library of Congress, "PREMIS Data Dictionary for Preservation Metadata, Version 2.2."

[58] Ibid.

[59] Allyson Carlyle, "FRBR and the Bibliographic Universe, or, How to Read FRBR as a Model" (presented at the ALA Annual Conference, Orlando, FL, 2004), http://www.ala.org/alcts/sites/ala.org.alcts/files/content/events/pastala/annual/04/Carlyle.pdf.

[60] Allen Renear et al., "An XML Document Corresponds to Which FRBR Group 1 Entity?," in *Proceedings of Extreme Markup Languages 2003* (presented at the Extreme Markup Languages, Montreal, Quebec, 2003), https://www.ideals.illinois.edu/handle/2142/11885.

[61] Michael Buckland, "What Is a Digital Document?," *Document Numerique (Paris)* 2, no. 2 (1998): 221–230.

[62] Ingbert R. Floyd and Allen H. Renear, "What Exactly Is an Item in the Digital World?," *Proceedings of the American Society for Information Science and Technology* 44, no. 1 (2007): 1–7, doi:10.1002/meet.1450440374.

[63] Harry Halpin, "The Principle of Self-description: Identity through Linking," *Proceedings of the 1st IRSW 2008* (2008), http://ceur-ws.org/Vol-422/irsw2008-submission-13.pdf.

[64] Caplan, *Understanding PREMIS*; Giuseppe Vitiello, "Identifiers and Identification Systems," *D-Lib Magazine* 10, no. 1 (January 2004), doi:10.1045/january2004-vitiello.

[65] Library of Congress, "PREMIS Data Dictionary for Preservation Metadata, Version 2.2."

[66] Caplan, *Understanding PREMIS*.

[67] Library of Congress, "PREMIS Data Dictionary for Preservation Metadata, Version 2.2."

[68] International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records*.

[69] ICOM/CIDOC CRM SIG, *Definition of the CIDOC Conceptual Reference Model*.

[70] Wynholds, "Linking to Scientific Data."

[71] Tom Pollard and J Wilkinson, "Making Datasets Visible and Accessible: DataCite's First Summer Meeting," *Ariadne* 64 (2010), http://www.ariadne.ac.uk/issue64/datacite-2010-rpt.

[72] Simmhan, Plale, and Gannon, "A Survey of Data Provenance in E-science"; Besiki Stvilia et al., "A Framework for Information Quality Assessment," *JASIST* 58 (2007): 1720–1733; "PROV Model Primer," 2013, http://www.w3.org/TR/prov-primer/.

[73] International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records*.

[74] ICOM/CIDOC CRM SIG, *Definition of the CIDOC Conceptual Reference Model*.

[75] International Organization for Standardization, *Information and Documentation - International Standard Name Identifier (ISNI)*, 2012, http://www.iso.org/iso/catalogue_detail?csnumber=44292.

[76] Data & GIS Lab, "GIS Across the Disciplines," 2013, http://libguides.ucsd.edu/content.php?pid=42741&sid=1825758.

[77] International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records*.

[78] ICOM/CIDOC CRM SIG, *Definition of the CIDOC Conceptual Reference Model*.

[79] GeoNames, "About GeoNames."

[80] ICOM/CIDOC CRM SIG, *Definition of the CIDOC Conceptual Reference Model*.

[81] Library of Congress, "PREMIS Data Dictionary for Preservation Metadata, Version 2.2."

[82] ICOM/CIDOC CRM SIG, *Definition of the CIDOC Conceptual Reference Model*.

[83] International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records*.

[84] "PROV Model Primer."

[85] Google Developers, "Google Schema," 2012, https://developers.google.com/public-data/docs/schema/dspl9.

[86] International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records*.

[87] Ibid.

[88] NISO/NFAIS, *Recommended Practices for Online Supplemental Journal Article Materials*.

[89] Jian Qin, Alex Ball, and Jane Greenberg, "Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data," in *Proceedings of International Conference on Dublin Core and Metadata Applications* (Kuching, Sarawak, Malaysis, 2012).

[90] DataUp, "DataUp," n.d., http://dataup.cdlib.org/.

[91] Toby Green, "We Need Publishing Standards for Datasets and Data Tables" (OECD Publishing, 2009), http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2010/wp.8.e.pdf.

[92] Altman and King, "A Proposed Standard for the Scholarly Citation of Quantitative Data."

[93] Duerr et al., "On the Utility of Identification Schemes for Digital Earth Science Data."

[94] Altman and King, "A Proposed Standard for the Scholarly Citation of Quantitative Data."

[95] Lee and Stvilia, "Identifier Schemas and Research Data."

[96] Besiki Stvilia et al., "Composition of Scientific Teams and Publication Productivity at a National Science Lab," *J. Am. Soc. Inf. Sci. Technol.* 62, no. 2 (February 2011): 270–283, doi:10.1002/asi.21464; Charles C. Hinnant et al., "Author-team Diversity and the Impact of Scientific Publications: Evidence from Physics Research at a National Science Lab," *Library & Information Science Research* 34, no. 4 (October 2012): 249–257, doi:10.1016/j.lisr.2012.03.001.

[97] National Institutes of Health, "NIH Data Sharing Policy and Implementation Guidance (NIH Publication No. 03-05-2003)"; National Science Foundation, "Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans (NSF 10-077)."

[98] DataCite, "DataCite," accessed November 6, 2012, http://datacite.org/.

[99] Mercè Crosas, "The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data," *D-Lib Magazine* 17, no. 1/2 (January 2011), doi:10.1045/january2011-crosas.

[100] William Michener et al., "DataONE: Data Observation Network for Earth — Preserving Data and Enabling Innovation in the Biological and Environmental Sciences," *D-Lib Magazine* 17, no. 1/2 (January 2011), doi:10.1045/january2011-michener.

[101] T Heath, "Linked Data," n.d., http://linkeddata.org/home.

[102] Christian Bizer, Tom Heath, and Tim Berners-Lee, "Linked Data - The Story so Far," *International Journal on Semantic Web and Information Systems* 5, no. 3 (2009): 1–22, doi:10.4018/jswis.2009081901.

[103] Tim Berners-Lee, "Linked Data" (W3C, 2006), http://www.w3.org/DesignIssues/LinkedData.html.

[104] Ibid.

[105] World Wide Web Consortium (W3C), "Resource Description Framework (RDF): Concepts and Abstract Syntax," 2004, http://www.w3.org/TR/rdf-concepts/.

[106] Chris Bizer, Richard Cyganiak, and Tom Heath, "How to Publish Linked Data on the Web," 2007, http://www4.wiwiss.fu-berlin.de/bizer/pub/linkeddatatutorial/.

[107] Stvilia et al., "Studying the Data Practices of a Scientific Community."

[108] Aalbersberg and Kähler, "Supporting Science through the Interoperability of Data and Articles."

[109] D Abbott, "Annotation," *DCC Briefing Papers: Introduction to Curation* (2008), http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/annotation.

[110] Wu, Stvilia, and Lee, "Authority Control for Scientific Data."

[111] K. D. Pruitt et al., "NCBI Reference Sequences: Current Status, Policy and New Initiatives," *Nucleic Acids Research* 37, no. Database (January 1, 2009): D32–D36, doi:10.1093/nar/gkn721.

[112] W3C, "Open Annotation Data Model," 2013, http://www.w3.org/ns/oa.

[113] Akhondi, Kors, and Muresan, "Consistency of Systematic Chemical Identifiers Within and Between Small-molecule Databases"; Altman and King, "A Proposed Standard for the Scholarly Citation of Quantitative Data"; Berners-Lee, "Cool URIs Don't Change"; Amy Brand, Frank Daly, and Barbara Meyers, *Metadata Demystified* (Sheridan and NISO Press, 2003), http://www.niso.org/standards/resources/Metadata_Demystified.pdf; Sarah Callaghan et al., "Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres," *International Journal of Digital Curation* 7, no. 1 (October 3, 2012): 107–113, doi:10.2218/ijdc.v7i1.218; Andrew Clark, "Anonymising Research Data" (ESRC National Centre for Research Methods, 2006), http://eprints.ncrm.ac.uk/480/1/0706_anonymising_research_data.pdf; Clark, Martin, and Liefeld, "Globally Distributed Object Identification for Biological Knowledgebases"; Crosas, "The Dataverse Network®"; Duerr et al., "On the Utility of Identification Schemes for Digital Earth Science Data"; N. Juty, N. Le Novère, and C. Laibe, "Identifiers.org and MIRIAM Registry: Community Resources to Provide Persistent Identification," *Nucleic Acids Research* 40, no. D1 (December 2, 2011): D580–D586, doi:10.1093/nar/gkr1097; Lee and Stvilia, "Identifier Schemas and Research Data"; Michener et al., "DataONE"; NISO/NFAIS, *Recommended Practices for Online Supplemental Journal Article Materials*; Carl Lagoze et al., "Specification and XML Schema for the OAI Identifier Format," March 9, 2006, http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm; Paskin, "Digital Object Identifier (DOI®) System"; Pepler and O'Neil, *Preservation Intent and Collection Identifiers*; Tonkin, "Persistent Identifiers: Considering the Options"; Vitiello, "Identifiers and Identification

Systems"; Toblas Weigel et al., "A Framework for Extended Persistent Identification of Scientific Assets," *Data Science Journal* 12 (2013): 10–22.

[114] R Wang and D Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers" 12, no. 4, Journal of Management Information Systems (1996): 5–35.

[115] M Eppler, *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes* (Berlin, Germany: Springer-Verlag, 2003); Stvilia et al., "A Framework for Information Quality Assessment."

[116] Berners-Lee, "Cool URIs Don't Change."

[117] NISO/NFAIS, *Recommended Practices for Online Supplemental Journal Article Materials*.

[118] Tonkin, "Persistent Identifiers: Considering the Options."

[119] Erway, *Lasting Impact*.

[120] Clark, Martin, and Liefeld, "Globally Distributed Object Identification for Biological Knowledgebases."

[121] Tim Berners-Lee, "Message on Www-tag@w3.org List," 2003, http://lists.w3.org/Archives/Public/www-tag/2003Jul/0022.html.; Tim Berners-Lee, "Message to Www-tag@w3.org List," 2003, http://lists.w3.org/Archives/Public/www-tag/2003Jul/0022.html.

[122] Harry Halpin, "Sense and Reference on the Web," *Minds and Machines* 21, no. 2 (2011): 153–178.

[123] P Hayes, "RDF Semantics," 2004, http://www.w3.org/TR/rdf-mt/.

[124] D Abbott, "Interoperability," *DCC Briefing Papers: Introduction to Curation* (2009), http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/interoperability.

[125] Paskin, "Identifier Interoperability."

[126] Gordon Dunsire, "Distinguishing Content from Carrier," *D-Lib Magazine* 13, no. 1/2 (January 2007), doi:10.1045/january2007-dunsire; Paskin, "Identifier Interoperability."

[127] Gustavo Pabón et al., "Linked Open Data Technologies for Publication of Census Microdata," *Journal of the American Society for Information Science and Technology* 64, no. 9 (2013): 1802–1814, doi:10.1002/asi.22876.

[128] Paskin, "Digital Object Identifier (DOI®) System."

[129] Lee and Stvilia, "Identifier Schemas and Research Data."

[130] Michener et al., "DataONE."

[131] Lee and Stvilia, "Identifier Schemas and Research Data."

[132] Buckland, "What Is a Digital Document?"; Carlyle, "FRBR and the Bibliographic Universe, or, How to Read FRBR as a Model"; Floyd and Renear, "What Exactly Is an Item in the Digital World?"; Renear et al., "An XML Document Corresponds to Which FRBR Group 1 Entity?".

[133] European Library Automation Group (ELAG), "Workshop on FRBR and Identifiers"; Halpin, "The Principle of Self-description: Identity through Linking"; Vitiello, "Identifiers and Identification Systems."

[134] Clark, Martin, and Liefeld, "Globally Distributed Object Identification for Biological Knowledgebases."

[135] Altman and King, "A Proposed Standard for the Scholarly Citation of Quantitative Data"; Michener et al., "DataONE."

[136] American Psychological Association (APA), *Publication Manual of the APA*, 6th ed. (Washington, D.C., 2010).

[137] Chemical Abstracts Service (CAS), "CAS Information Use Policies," 2012, http://www.cas.org/legal/infopolicy; Clark, Martin, and Liefeld, "Globally Distributed Object Identification for Biological Knowledgebases"; GeoNames, "About GeoNames"; Norman Paskin, "Digital Object Identifiers for Scientific Data," *Data Science Journal* 4 (2005), http://www.doi.org/topics/041110CODATAarticleDOI.pdf; Pruitt et al., "NCBI Reference Sequences."

[138] Paskin, "Identifier Interoperability."

[139] Altman and King, "A Proposed Standard for the Scholarly Citation of Quantitative Data."

[140] Michener et al., "DataONE."

[141] Duerr et al., "On the Utility of Identification Schemes for Digital Earth Science Data."

[142] Altman and King, "A Proposed Standard for the Scholarly Citation of Quantitative Data."

[143] Vitiello, "Identifiers and Identification Systems."

[144] LeBoeuf, "Identifying 'Textual Works': ISTC: Controversy and Potential."

[145] European Library Automation Group (ELAG), "Workshop on FRBR and Identifiers."

[146] Simmhan, Plale, and Gannon, "A Survey of Data Provenance in E-science"; "PROV Model Primer."

[147] Stvilia et al., "Studying the Data Practices of a Scientific Community."